1. Instrumental Variables

EXAMPLE 15.2 ESTIMATING THE RETURN TO EDUCATION FOR MEN

We now use WAGE2.RAW to estimate the return to education for men. We use the variable *sibs* (number of siblings) as an instrument for *educ*. These are negatively correlated, as we can verify from a simple regression:

$$\widehat{educ} = 14.14 - .228\, sibs$$
$$(.11)\quad (.030)$$
$$n = 935,\ R^2 = .057.$$

This equation implies that every sibling is associated with, on average, about .23 less of a year of education. If we assume that *sibs* is uncorrelated with the error term in (15.14), then the IV estimator is consistent. Estimating equation (15.14) using *sibs* as an IV for *educ* gives

$$\widehat{\log(wage)} = 5.13 + .122\, educ$$
$$(.36)\quad (.026)$$
$$n = 935.$$

(The *R*-squared is computed to be negative, so we do not report it. A discussion of *R*-squared in the context of IV estimation follows.) For comparison, the OLS estimate of β_1 is .059 with a standard error of .006. Unlike in the previous example, the IV estimate is now much higher than the OLS estimate. While we do not know whether the difference is statistically significant, this does not mesh with the omitted ability bias from OLS. It could be that *sibs* is also correlated with ability: more siblings means, on average, less parental attention, which could result in lower ability. Another interpretation is that the OLS estimator is biased toward zero because of measurement error in *educ*. This is not entirely convincing because, as we discussed in Section 9.3, *educ* is unlikely to satisfy the classical errors-in-variables model.

Nota: Problema C1 a seguir se refere ao exemplo 15.2.

C1 Use the data in WAGE2.RAW for this exercise.

(i) In Example 15.2, if *sibs* is used as an instrument for *educ*, the IV estimate of the return to education is .122. To convince yourself that using *sibs* as an IV for *educ* is *not* the same as just plugging *sibs* in for *educ* and running an OLS regression, run the regression of log(*wage*) on *sibs* and explain your findings.

(ii) The variable *brthord* is birth order (*brthord* is one for a first-born child, two for a second-born child, and so on). Explain why *educ* and *brthord* might be negatively correlated. Regress *educ* on *brthord* to determine whether there is a statistically significant negative correlation.

(iii) Use *brthord* as an IV for *educ* in equation (15.1). Report and interpret the results.

(iv) Now, suppose that we include number of siblings as an explanatory variable in the wage equation; this controls for family background, to some extent:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 sibs + u.$$

Suppose that we want to use *brthord* as an IV for *educ*, assuming that *sibs* is exogenous. The reduced form for *educ* is

$$educ = \pi_0 + \pi_1 sibs + \pi_2 brthord + v.$$

State and test the identification assumption.

(v) Estimate the equation from part (iv) using *brthord* as an IV for *educ* (and *sibs* as its own IV). Comment on the standard errors for $\hat{\beta}_{educ}$ and $\hat{\beta}_{sibs}$.

(vi) Using the fitted values from part (iv), $\widehat{educ}$, compute the correlation between $\widehat{educ}$ and *sibs*. Use this result to explain your findings from part (v).

EXAMPLE 15.4 USING COLLEGE PROXIMITY AS AN IV FOR EDUCATION

Card (1995) used wage and education data for a sample of men in 1976 to estimate the return to education. He used a dummy variable for whether someone grew up near a four-year college (*nearc4*) as an instrumental variable for education. In a log(*wage*) equation, he included other standard controls: experience, a black dummy variable, dummy variables for living in an SMSA and living in the South, and a full set of regional dummy variables and an SMSA dummy for where the man was living in 1966. In order for *nearc4* to be a valid instrument, it must be uncorrelated with the error term in the wage equation—we assume this—and it must be partially correlated with *educ*. To check the latter requirement, we regress *educ* on *nearc4* and all of the exogenous variables appearing in the equation. (That is, we estimate the reduced form for *educ*.) Using the data in CARD.RAW, we obtain, in condensed form,

$$\begin{aligned} educ &= 16.64 + .320\, nearc4 - .413\, exper + \ldots \\ &\qquad (.24)\ \ (.088)\qquad\quad (.034) \\ n &= 3{,}010,\ R^2 = .477. \end{aligned}$$

We are interested in the coefficient and *t* statistic on *nearc4*. The coefficient implies that in 1976, other things being fixed (experience, race, region, and so on), people who lived near a college in 1966 had, on average, about one-third of a year more education than those who did not grow up near a college. The *t* statistic on *nearc4* is 3.64, which gives a *p*-value that is zero in the first three decimals. Therefore, if *nearc4* is uncorrelated with unobserved factors in the error term, we can use *nearc4* as an IV for *educ*.

The OLS and IV estimates are given in Table 15.1. Interestingly, the IV estimate of the return to education is almost twice as large as the OLS estimate, but the standard error of the IV estimate is over 18 times larger than the OLS standard error. The 95% confidence interval for the IV estimate is between .024 and .239, which is a very wide range. The presence of larger confidence intervals is a price we must pay to get a consistent estimator of the return to education when we think *educ* is endogenous.

TABLE 15.1 Dependent Variable: log(*wage*)

Explanatory Variables	OLS	IV
educ	.075 (.003)	.132 (.055)
exper	.085 (.007)	.108 (.024)
$exper^2$	−.0023 (.0003)	−.0023 (.0003)
black	−.199 (.018)	−.147 (.054)
smsa	.136 (.020)	.112 (.032)
south	−.148 (.026)	−.145 (.027)
Observations *R*-squared	3,010 .300	3,010 .238
Other controls: *smsa66, reg662, …, reg669*		

As discussed earlier, we should not make anything of the smaller *R*-squared in the IV estimation: by definition, the OLS *R*-squared will always be larger because OLS minimizes the sum of squared residuals.

Nota: Problema C5 a seguir se refere ao exemplo 15.4.

C5 Use the data in CARD.RAW for this exercise.

(i) In Table 15.1, the difference between the IV and OLS estimates of the return to education is economically important. Obtain the reduced form residuals, $\hat{v}_2$, from the reduced form regression *educ* on *nearc*4, *exper*, *exper*^{2}, *black*, *smsa*, *south*, *smsa*66, *reg*662, ..., *reg*669—see Table15.1. Use these to test whether *educ* is exogenous; that is, determine if the difference between OLS and IV is *statistically* significant.

(ii) Estimate the equation by 2SLS, adding *nearc2* as an instrument. Does the coefficient on *educ* change much?

(iii) Test the single overidentifying restriction from part (ii).

2. Simultaneous Equations

2 Let *corn* denote per capita consumption of corn in bushels at the county level, let *price* be the price per bushel of corn, let *income* denote per capita county income, and let *rainfall* be inches of rainfall during the last corn-growing season. The following simultaneous equations model imposes the equilibrium condition that supply equals demand:

$$corn = \alpha_1 price + \beta_1 income + u_1$$
$$corn = \alpha_2 price + \beta_2 rainfall + \gamma_2 rainfall^2 + u_2.$$

Which is the supply equation, and which is the demand equation? Explain.

4 Suppose that annual earnings and alcohol consumption are determined by the SEM

$$\log(earnings) = \beta_0 + \beta_1 alcohol + \beta_2 educ + u_1$$
$$alcohol = \gamma_0 + \gamma_1 \log(earnings) + \gamma_2 educ + \gamma_3 \log(price) + u_2,$$

where *price* is a local price index for alcohol, which includes state and local taxes. Assume that *educ* and *price* are exogenous. If β_1, β_2, γ_1, γ_2, and γ_3 are all different from zero, which equation is identified? How would you estimate that equation?

C1 Use SMOKE.RAW for this exercise.

(i) A model to estimate the effects of smoking on annual income (perhaps through lost work days due to illness, or productivity effects) is

$$\log(income) = \beta_0 + \beta_1 cigs + \beta_2 educ + \beta_3 age + \beta_4 age^2 + u_1,$$

where *cigs* is number of cigarettes smoked per day, on average. How do you interpret β_1?

(ii) To reflect the fact that cigarette consumption might be jointly determined with income, a demand for cigarettes equation is

$$cigs = \gamma_0 + \gamma_1 \log(income) + \gamma_2 educ + \gamma_3 age + \gamma_4 age^2 + \gamma_5 \log(cigpric) + \gamma_6 restaurn + u_2,$$

where *cigpric* is the price of a pack of cigarettes (in cents), and *restaurn* is a binary variable equal to unity if the person lives in a state with restaurant smoking restrictions. Assuming these are exogenous to the individual, what signs would you expect for γ_5 and γ_6?

(iii) Under what assumption is the income equation from part (i) identified?

(iv) Estimate the income equation by OLS and discuss the estimate of β_1.

(v) Estimate the reduced form for *cigs*. (Recall that this entails regressing *cigs* on all exogenous variables.) Are log(*cigpric*) and *restaurn* significant in the reduced form?

(vi) Now, estimate the income equation by 2SLS. Discuss how the estimate of β_1 compares with the OLS estimate.

(vii) Do you think that cigarette prices and restaurant smoking restrictions are exogenous in the income equation?